

Within-PSU Sort Research to Reduce Variances for the Survey of Income and Program Participation (SIPP)¹

Sarah T. McMillan (Sarah M. Tekansik)

U.S. Census Bureau

sarah.tekansik.mcmillan@census.gov

Stephen P. Mack

U.S. Census Bureau

stephen.p.mack@census.gov

I. Introduction

Every ten years, following the Decennial Census, the Census Bureau demographic surveys redesign their sample selection process to determine ways to improve each survey's efficiency. Most of the demographic surveys, including the Survey of Income and Program Participation (SIPP), have a two-stage sampling process. The first stage of sampling consists of selecting a stratified random sample of groups of counties (primary sampling units or PSUs). The second stage consists of selecting a systematic sample of households from the selected PSUs. The main focus of this research project included determining the variables to use for sorting households before the systematic sample in the second stage of sampling. The goal is to find a set of variables that produces a sort scheme which minimizes the within-PSU variance. Minimizing this variance can improve the survey's efficiency.

For the 2010 Sample Redesign, the SIPP is reevaluating two different aspects for the sort. The first is the type of variables, whether they should be geographic, demographic, social, financial or some combination. The second is the sorting scheme to be used, which includes what order and what source the data will come from.

II. Background

Survey of Income and Program Participation

The SIPP is a longitudinal survey developed to improve the quality of household income information and to collect detailed information on eligibility and participation in government programs. The SIPP has many key estimates, including: average monthly rates of poverty and program participation, number of people receiving social security or retirement income, estimates of total household income and estimates of health insurance coverage [3]. Each household sample for the SIPP is called a panel. For each panel in the previous redesign, sample households were interviewed every four months for three to four years. With the new redesign, each panel will be interviewed once each year for three years.

Sampling Frame

In previous redesigns, the Decennial Census file served as the sampling frame for all demographic surveys and each survey would choose ten years worth of sample, to be used throughout the decade. For this redesign, the sampling frame will be the Master Address File (MAF) and sample will be chosen on an annual basis. Each survey's sample is chosen independently and removed to ensure that the remaining frame is an unbiased universe. This ensures that each sample is representative of the population. The population represented in the SIPP is the civilian noninstitutionalized population living in the United States.

Sample Design

Sample for the SIPP is chosen using a two stage sampling process. The first stage of sampling consists of selecting a stratified random sample of groups of counties (primary sampling units or PSUs). The second stage consists of selecting a systematic sample of households from the selected PSUs. The sample size for the 2008 Panel was about

¹ This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress). Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

65,000 households, which yielded about 42,000 interviews. The new SIPP Panel will have approximately the same sample size.

An oversample of the low income population has been a part of the past two redesigns of the SIPP samples. Oversampling occurs when certain groups or units are sampled with higher probabilities than other groups or units. This allows analysts more low income cases to strengthen their analyses of government programs or other characteristics of the low income population. Oversampling is done by separating housing units in each PSU into two strata: low and high income. In the 2000 design, Census long form data was used to determine the correct strata to place each household. The two strata are then sampled at different rates, higher in the low income and lower in the high income to achieve the desired sample size.

Previous Methods and Research on Sorting Schemes

For the 1990 design, variables were chosen from the decennial census based on their correlations with key estimates (which included total population 0+ below poverty level, total blacks 16+ in the civilian labor force, and total female-headed households), their ability to reduce variances, and their stability over time. The research was done at the block level, but then the variables were used to sort and stratify both blocks and units. Race, tenure, median rent, property value, and CBUR (a classification variable indicating whether the area was urban or rural) were found to be the best sort and stratification variables [1]. These demographic and financial variables were used along with some geographic variables state, county, block, and unit ID.

In the 2000 design, there was limited time and money to conduct research for within-PSU sampling. The sorting scheme used was similar to what was implemented in the previous design, including variables such as race, tenure, and CBUR from the Decennial Census short form. The low income variable used for oversampling was created using data from the Decennial Census Long Form, and was an estimate of the household's probability of being low income based on the data from the long form respondents in that area. These, along with geography variables, formed the sorting scheme for this design.

For the 2010 sample redesign, many things about the sample design are changing, so it was imperative that research be conducted to determine the best sorting scheme.

III. Methodology

This research involved two main parts: determining possible sort variables and determining the combination of variables that would produce a sort scheme that will minimize the within-PSU variance. The latter included determining the source the variables should come from. To determine the best variables and schemes for consideration, the research included two analyses: 1) identifying variables correlated to our key estimates and 2) selecting samples using various sort schemes and calculating the variance for each of the key estimates. To determine the best source of information, we looked at whether recent data summarized to block, tract, or county levels from the ACS was better at predicting our key estimates than unit level information collected from the 2000 Decennial Census.

Data Used for Research

The American Community Survey (ACS) served as one of the sources of data for this research. The ACS collects data similar to data that was collected on the Decennial Census Long form, but instead of data collected once per decade like the long form, the ACS is collected on a rolling basis each year. Each household selected for sample is asked basic demographic questions like age, sex, and race but also social and financial questions about income, health insurance, education, and disabilities pertaining to the previous twelve months. ACS interviews are conducted throughout the year, so the data is not measuring one point in time. The benefit of the ACS data over the Decennial Census Long Form data is that we can get more recent information, but a single year of ACS data only provides about 20% of the sample formerly available from the Decennial Census long form data. Five years of ACS can be combined to represent about 1 in 8 households [7].

The ACS variables listed in Table 1 were considered for the sorting scheme. Each variable either directly corresponds to an ACS question or is created from a combination of ACS questions. The low income variable used to create the oversampling strata in the previous redesign was created using ACS data (PPOV) for this redesign.

| Table 1 | |
|--|---------------|
| Variable Description | Variable Name |
| Probability of low income | PPOV |
| Indicator of nonwhite householder | NWHT |
| Household tenure | TEN |
| Indicator of renter household | RENT |
| Number of persons in household | NPHU |
| Inflation adjusted household income | HINC |
| Number of earners per occupied household | EARN |
| Highest level of schooling completed for householder | SCHL |
| Number of rooms per household | NRMS |
| Number of beds per household | NBDS |
| Year built | YRBLT |
| Property value | PRVAL |
| Monthly housing cost | MHC |
| Number of persons per room | NPLRM |
| Number of unemployed householders | UNEMP |
| Indicator of single female head of household | FHH |

The 2000 Decennial Census was the other source of data used in research. The short form included questions on tenure, name, sex, age, relationship to householder, race, and Hispanic origin [6]. The long form was sent to about one in six households and included additional questions on social characteristics (marital status, education), economic characteristics (labor force status, income), physical characteristics of housing (number of rooms, year built), and financial characteristics of housing (value of home, monthly rent) [6]. Data from the long form was used for parts of this research, but is not included in this paper. The Decennial variables included in Table 2 were those we considered for research.

| Table 2 | |
|---|---------------|
| Variable Description | Variable Name |
| Number of people in household | CNPHU |
| Indicator of renter household | CRENT |
| Probability of poverty (taken from Census long forms) | CPPOV |
| Indicator of whether householder is a minority | CNWHT |
| Indicator of female headed household | CFHH |

Creating Key Estimates

The SIPP was undertaken to collect more detailed information on income, with a particular interest in persons with low income and receiving program assistance. There are several other important estimates for SIPP, but for our analysis, we considered these the key estimates. For the correlation analysis, we created four key estimates using SIPP 2008 data to calculate state level proportions or means. The first and second deal with proportions of households in poverty, the first being those below 100% of the poverty threshold (POV) and the second was the proportion of households below 150% of the poverty threshold (POV150). The third key estimate was the proportion of households that participated in any welfare assistance program (PROGPT), including WIC, food stamps, and supplemental security income. The last key estimate was mean household income (INC).

For our variance analysis, we needed unit level information on poverty and program participation for all eligible households, so we created ACS estimates similar to our SIPP key estimates. The ACS determines poverty based on the primary family's (not necessarily the entire household) income to the poverty threshold of the family. For the SIPP, the entire household income is compared to the household's poverty threshold. Bearing this in mind, we created two different estimates of poverty using variables on the ACS. The first uses the family poverty indicator variable from ACS. The second compares the household's annual income to 150% of the household's poverty threshold (recalculated to take into account all persons in household). The program participation ACS estimate

calculated was similar to the SIPP estimate of the proportion of households that participated in any welfare program and the mean income estimate was calculated using ACS household income.

Creating Sort Variables

For the correlation analysis, we summarized variables from the ACS and the Decennial Short Form at the state level to compare to SIPP state level key estimates. To be able to perform a sort data is needed for each unit on the frame. Because the ACS is a survey and only has data only for sampled and interviewed units, we needed to calculate summary measures at the different levels of geography in order to give a value to every unit on the frame. We created proportions and means for blocks, tracts, and counties. The geography level is not based on population size, so some blocks, tracts, and counties have a large number of households while others have only a few households. When creating block level proportions using ACS data, some proportions or means might be calculated using only one or two unweighted occupied households. To counter this problem, we calculated proportions and means using higher and higher levels of geography until the proportion or mean was based on at least twenty occupied households. For example, if the block did not have enough observations then we checked the tract. If the tract didn't have twenty occupied households, we went to the county level.

Decennial Short Form variables are available at the unit level, so there is no need to summarize the data at different geographic levels.

Correlation Analysis

Choosing variables to use in sorting that are related to the key estimates can lead to a reduction in the within-PSU variance, since we are trying to group like units together. The correlation between two variables tells us the strength of the linear relationship between them or how the variables are related. Therefore, our first step in this research was to undertake a correlation analysis to determine possible variables to use in sorting that are correlated to our key estimates.

We used SAS PROC CORR to obtain Pearson Correlation Coefficients for each of the possible sort variables from ACS and Decennial with our key SIPP estimates. We included the variables with the highest coefficients with our key estimates in sorting schemes for the variance analysis.

Variance Analysis

From the correlation analysis, we determined which variables were best at predicting our key estimates, but we still needed to determine the order and source of the variables for a sorting scheme that would minimize the variance for our key estimates.

To be able to evaluate the impact of the sort on minimizing variance, we needed to simulate sample selection from a frame that could be sorted by various sorting schemes. Previously, this research did not select different samples for the various sorting schemes. SIPP sample cases were sorted according to the various schemes and then used to calculate variances. The SIPP sample had already been chosen according to a sorting scheme, so changing the sort was not affecting which units would be selected. This did not seem to be evaluating the different schemes, since sorting sample units on a frame according to different schemes would change the units that would be selected into sample and used to calculate variance estimates. Therefore, we decided to start from a frame, instead of sampled units, and select sample. This allowed us to select different households into sample. For the frame, we decided to use the ACS five-year file. Even though the ACS has its own within-PSU sort, the sample is large enough to use as a frame for selecting samples. The Decennial Census was considered, but unit level information similar to our key estimates was not available to use for calculating variances.

To evaluate each of the schemes, we first sorted the frame according to the scheme. Next, we selected a sample using PROC SURVEYSELECT in SAS that was approximately the size of the 2008 SIPP Panel. Then the variances for our key estimates were calculated and compared to the variance of a random sort.

The goal of sorting is to group units together that have similar values of poverty, program participation, and income. We wanted to choose a method of calculating variance that would take into account the differences between successive units. If units with similar values of the key variable are sorted together, then the variance will be smaller than the variance using a random or geographic sort. Therefore, to calculate an estimate of variance for our key estimates, we used the method of successive differences in the following form

$$V = \sum_c \frac{1}{2N_c(N_c - 1)} \sum_i^{N_c-1} w_i (y_{i+1} - y_i)^2$$

where C is the total number of counties, N_c is the number of units in county c , i indicates the sort order (the i^{th} unit in the order), y_i represents the variable of interest of the i^{th} unit, w_i is the weight associated with i^{th} unit, and V is the variance for the estimate of interest.

To determine the best sorting scheme, we took the ratio of the variance calculated from the test sorting scheme to the variance calculated from no sorting scheme. To impose no sorting scheme we created a random number for each unit and sorted by that random number. To make the results easier to understand, we subtracted the ratio variance from one and called it RV . This creates a value similar to a design effect and RV tells us the percent reduction in variance using the new sorting scheme compared to having no sorting scheme. The formula is given below.

$$RV = 1 - \frac{V_{\text{new sort}}}{V_{\text{random sort}}}$$

IV. Results

Correlation Results

Table 3 provides the Pearson correlation coefficients of ACS possible sort variables with the SIPP key estimates. The variables used in previous redesigns, including indicator of renter and indicator of nonwhite household, did not have high correlations as expected. For these variables, most coefficients were not significantly different from zero. Some of the variables that did have high correlations and were significant in all key estimates were probability of low income (PPOV), household income (HINC), highest level schooling attained (SCHL), number of earners in the household (NEARN), and monthly housing cost (MHC).

| Variable | POV | POV150 | PROGPT | INC |
|----------|--------|--------|--------|--------|
| PPOV | 0.58* | 0.59* | 0.50* | -0.48* |
| NWHT | 0.20 | 0.06 | 0.29* | 0.31* |
| TEN | 0.23 | 0.31 | 0.37* | -0.50 |
| RENT | -0.05 | -0.13 | 0.09 | 0.35 |
| HINC | -0.54* | -0.65* | -0.43* | 0.89* |
| SCHL | -0.68* | -0.69* | -0.58* | 0.71* |
| NROOM | -0.18 | -0.12 | -0.41* | -0.07 |
| NBED | -0.07 | 0.01 | -0.18 | -0.22 |
| NEARN | -0.79* | -0.77* | -0.63* | 0.57* |
| MHC | -0.53* | -0.62* | -0.39* | 0.86* |
| NPPLRM | -0.12 | -0.18 | 0.19 | 0.37* |
| FHH | 0.41* | 0.28* | 0.34* | 0.08 |

We also performed the correlations for Decennial variables with the SIPP key estimates. Only the coefficients for the probability of low income (CPPOV) were significant for all key estimates.

² * Indicates that the coefficient is significantly different from zero at the 0.05 level.

| Table 4 ³ | | | | |
|----------------------|-------|--------|--------|--------|
| Variable | POV | POV150 | PROGPT | INC |
| CPPOV | 0.79* | 0.83* | 0.73* | -0.67* |
| CNWHT | 0.44* | 0.28* | 0.38* | 0.14 |
| CTEN | -0.01 | -0.09 | 0.05 | 0.33* |
| CRENT | -0.16 | -0.26* | 0.38* | 0.14 |
| CNPHU | -0.2 | -0.25* | -0.11 | 0.39* |
| CFHH | 0.2 | 0.07 | 0.16 | 0.34* |

Variance Results

Because the ACS served as our sample frame and the source for our key estimates, we first wanted to make sure the reductions in variance seen in ACS estimates would possibly lead to reductions in variance of SIPP key estimates. Therefore, we calculated the Pearson Correlation Coefficients for the ACS key estimates and the SIPP key estimates. We include these results in Table 5 with those in bold being within category comparisons. The correlations seem to support the conclusion that reductions in variance seen in this analysis would be similar in magnitude to those that could be seen for SIPP.

| Table 5 ⁴ | | | | |
|----------------------|--------------|--------------|--------------|--------------|
| Variable | POV | POV150 | PROGPT | INC |
| POV(ACS) | 0.82* | 0.84* | 0.73* | -0.70* |
| PROGPT (ACS) | 0.37* | 0.35* | 0.63* | -0.15 |
| INC(ACS) | -0.54* | -0.65* | -0.43* | 0.89* |

We formulated different sorts using variables from the correlation analysis and geographic variables. Choosing different sorting schemes was difficult, because for some sets of variables PPOV was a better first variable, but then for others it was TEN. The sorts that had the largest reductions in variance or largest *RV* value were those that used the probability of low income, calculated from the ACS, and decennial variables. The ten-year old census data had larger reductions in variance than summarized ACS data, especially for estimates of program participation. Table 6 shows specific sorts and their *RV* values. The first four sorts have the largest *RV* values for particular key estimates (those indicated in bold). The remaining sorts had good overall reductions in variance for all of the key estimates.

| Table 6 | | | | |
|---|---------------|--------------|--------------|--------------|
| Sorting Scheme | RV values for | | | |
| | POV | POV150 | PROGPT | INC |
| CRENT, PPOV, CNWHT | 0.098 | 0.232 | 0.081 | 0.000 |
| CTEN, PPOV, SCHL | 0.096 | 0.242 | 0.052 | 0.067 |
| HINC, PPOV, SCHL, TEN, FEARN, FHH, NPPLPR | 0.079 | 0.223 | 0.049 | 0.177 |
| PPOV, CTEN, SCHL, FEARN | 0.069 | 0.219 | 0.057 | 0.130 |
| CTEN, PPOV, CNWT, NPHU | 0.086 | 0.229 | 0.068 | 0.210 |
| CTEN, PPOV, CNWT, CFHH | 0.079 | 0.224 | 0.077 | 0.083 |
| PPOV, FEARN, FHH, NPPLPR | 0.081 | 0.223 | 0.073 | 0.050 |

From Table 6, we see that there were some significant reductions in variance, especially for the estimate of those households below 150% of the poverty threshold. These results could be slightly inflated do to the similarity of variables used to create the sorting variables and key estimates. To interpret these results, a *RV* value of 0.232 means that overall, the sorting scheme using indicator of renter, probability of low income, and indicator of nonwhite household had a variance that was about 23% lower than the variance of the random sort. If we were to look at individual counties, we would see larger and smaller reductions in variance. From these results, we were

³ * Indicates that the coefficient is significantly different from zero at the 0.05 level.

⁴ * Indicates that the coefficient is significantly different from zero at the 0.05 level.

able to conclude that using the probability of poverty from the ACS and unit level data from the Decennial Short Form, even if older, led to the greatest reductions in variance for our key estimates.

V. Other Sort Research

While undertaking this research, several other issues arose in doing the sorting due to changes in the frame and the design. Most of these issues dealt with using the MAF as our frame including: how to handle the new growth units that get added to the MAF every six months, how to handle duplicate addresses, and how to use the numerous geographic variables in our sort. Only the issue regarding new growth has been fully resolved, so that issue is discussed below.

New Growth

For the 2000 design, new units built after the 2000 census were picked up from building permits and placed into a separate frame that was sorted geographically and sampled separately. The MAF receives new addresses from the post office delivery sequence file every six months. Using the MAF for our frame will allow us to have a current sampling frame, but we needed to determine how to sort these new units. There were two options for sorting: use only the geographic information on the new units to sort, or, in addition to the geographic variables, add summarized ACS data to the new units and sort.

Using our test file of ACS and decennial cases, we classified those ACS cases that did not match to the decennial, and that were built since 2000, as new growth. This would simulate the worst case scenario in terms of amount of new growth. We used a process similar to the variance analysis above, where a sort was performed on the frame, PROC SURVEYSELECT was used to select a sample, and, then, successive differences was used to calculate variances for our key estimates. We calculated the variance for sorting schemes of summarized ACS proportions and means for those units considered new growth against the variance of a purely geographic scheme. The ratio was then subtracted from one. We called this RV' . The RV' values for the best overall sorts are listed in Table 6.

| Table 6 | | | | |
|---------------------------------------|------------------|--------------|--------------|--------------|
| Best Overall Each Category Sort Order | RV' values for | | | |
| | POV | PROGPT | POV150 | INC |
| FHH, PPOV, SCHL, FEARN | 0.006 | -0.004 | 0.003 | -0.011 |
| HINC, PPOV, SCHL, FEARN | 0.002 | 0.007 | 0.002 | -0.02 |
| FEARN, SCHL, PPOV | 0.001 | 0.0001 | 0.003 | 0.009 |
| PPOV, FEARN, FHH, NPPLPR | 0.004 | -0.02 | -0.001 | 0.055 |

From the results, we can see that the effect on variance was negligible, even negative, for the key estimates. The largest reduction in variance was obtained for the estimates of income by sorting the new growth units by the probability of low income, number of earners in the household, and number of people per room. This sort had a variance that was about 5.5% lower than the variance using only geographic information for the new units. This sort also had a variance that was 2% higher than then geographic sort for the key estimate of program participation. Noting these contradicting results, we decided to sort new units added to the MAF by geographic variables.

VI. Conclusion

Limitations

One of the major limitations of this research was that we did not have a complete frame to be able to simulate our sorting and selection of samples. Using the ACS as our frame allowed us to simulate the process of sorting and selecting to see how different variables and levels of variables fared. The ACS has good coverage, but a sample of respondents falls short of the qualifications for a good frame.

Another limitation of our research is that only one method was used for each of the analyses. Comparing the variance from different methods could have given us a better idea of the actual impact on total variance sorting would have. Even though the correlations were high between the SIPP key estimates and the ACS key estimates, we do not know if the results will lead to actual reductions in variance for the SIPP. Also, we did not incorporate the complex sample design or the oversample into the samples chosen from the ACS. The procedure to create the

oversampling strata occurs after the initial sample selection and could have an affect on actual variance reductions seen for the SIPP.

The lack of external data sources to use to compare to our results was another limitation of our research. Our research would have been strengthened, if in the variance analysis we could have had independent sources for the sort variables and key estimates. The variance for the sorting schemes using census variables could be overstated, since when matching the ACS to the 2000 Census file, we are then restricting our sample to only those households that were present for the 2000 Census.

Lastly, the timing for our research did not allow us to look into the possibility of using administrative data like income data from the Internal Revenue Service (IRS) or recipient rates for food stamps or TANF. These could be valuable sources for sort and oversampling research in the future.

Future Samples

We determined that based on this research the best sort for the SIPP in the 2014 Panel would be to use the Decennial Census variables based on tenure, race of the householder, and number of people per household, along with the probability of low income variable created for oversampling from ACS data. With the implementation of annual sampling in this redesign, there is the possibility of changing the sort for the 2017 Panel. Future research could include looking into using administrative records for creating sort variables.

VII. References

- [1] Gorsak, M, Mansur K., Fenstermaker D, Petroni R. *Within-PSU Sort and Stratification Research to Improve Survey Efficiency*. (October 2, 1991) Retrieved August 2, 2010 from http://www2.census.gov/prod2/sipp/wp/SIPP_WP_155.pdf.
- [2] U.S. Census Bureau internal memorandum from Andrew Zbikowski for documentation dated July 20, 2011. *Research and Recommendations on 2010 Sample Redesign Within-PSU Sort Variables for the Current Population Survey*.
- [3] U.S. Census Bureau internal memorandum from Ruth Ann Killion to Howard Hogan dated September 25, 2009. *Analysis of Possible Benefits to the Re-engineered Survey of Income and Program Participation (SIPP) of Using the American Community Survey as a Frame*.
- [4] U.S. Census Bureau internal memorandum to Household Survey Sample Design WG 3.0 from Within-PSU Sampling Team WG3.8 dated August 4, 2010. *New Growth Research Results and Recommendation for the 2010 Demographic Surveys Sample Redesign*.
- [5] U.S. Census Bureau memorandum from Ruth Ann Killion to Jay Ryan dated December 8, 2009. *Consumer Expenditure Survey 2010 Redesign Research: Determining the Within-Primary Sampling Unit Stratification*.
- [6] *Website for the 2000 Decennial Census*. Introduction to Census 2000 Data Products. Retrieved September 9, 2011 from <http://www.census.gov/prod/2001pubs/mso-01icdp.pdf>.
- [7] *Website for American Community Survey*. Design and Methodology for the American Community Survey (2009). Retrieved August 16, 2010 from http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf.
- [8] *Website for Survey of Income and Program Participation*. SIPP Users' Guide. Retrieved September 9, 2011 from <http://www.census.gov/sipp/usrguide.html>.